Fall 2013

Patterns of intron loss and gain in Caenorhabditis

Gabrielle Giese
University of New Hampshire, Durham

PATTERNS OF INTRON LOSS AND GAIN IN *CAENORHABDITIS*

BY

Gabrielle Giese

B.A., Bennington College, 2005

THESIS

Submitted to the University of New Hampshire

in Partial Fulfillment of

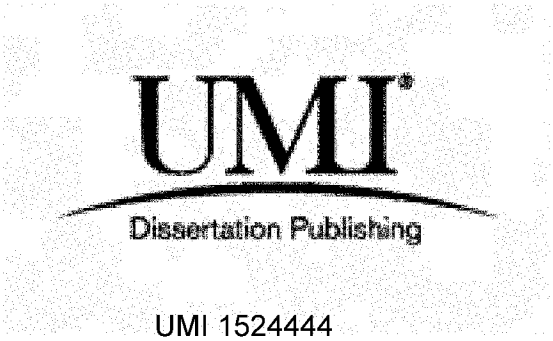
the Requirements for the Degree of

Masters of Science

in

Biochemistry

September, 2013



ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

This thesis has been examined and approved.

Thesis director, John J Collins
Associate Professor of Genetics

W. Kelley Thomas
Professor of Genetics

R. Daniel Bergeron
Professor of Computer Science

8/29/13
Date

ACKNOWLEDGEMENTS

There are no words in which I can express my gratitude for my parents for everything: their support both financially and emotionally, their indulgence, their love and friendship, and for being great human beings. You are still my heroes.

I would like to thank Kelley Thomas for helping me when I most had need of guidance. Without your help I would never have made it this far. I need to also thank Daniel Bergeron. Thank you for all your hard work and incredible patience.

I must also thank Feseha Abebe-Akele for his time and instruction in Perl. Thank you to everyone at the Hubbard Genome Center especially Krystalynne Morris and Feixia Chu. I would also like to express my appreciation for the people who wrote the original versions of several scripts I used in my thesis: Phill Hatcher, Sam Vohr and Way Sung.

And of course. I must thank my adviser John Collins. Thank you for the unique experience my graduate education has turned out to be. Thank you also for your one-of-a-kind wit and your ability to teach.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

PATTERNS OF INTRON LOSS AND GAIN IN *CAENORHABDITIS*

by

Gabrielle Giese

University of New Hampshire, September 2013

Introns, segments of genes that get spliced from the transcript before translation, are prevalent parts of many genomes and yet remain largely mysterious. Although their presence in the genome has been known for over thirty years, we still cannot answer the most fundamental questions about introns, such as where did they originate and, how are they gained and lost? In our most stringent dataset, we compared 137,453 intron positions in 6,257 pan-orthologs among four species of *Caenorhabditis* using a bioinformatics approach. While 82% of intron positions were conserved, we found a remarkable amount of intron variation. We also found evidence suggestion a role of both transcription-mediated processes as well as transposon activity in the gain of novel introns. These results suggest a more dynamic picture of intron gain and loss than previously thought and identify some mechanisms that may be responsible for intron gain.

INTRODUCTION

In the post-genomic era, with the entire human genome and the genomes of many other organisms sequenced and available to anyone with a computer, the mysteries of the genome are being revealed. Bioinformatics allows the storage and annotation of genomic information that can be easily accessed for further analysis. The result is a major shift in the way science is done. More and more, questions are being addressed on the computer, rather than the lab bench. This new frontier of science will be instrumental in revealing the architecture of the genome.

The nematode *Caenorhabditis elegans* was the first multicellular organism to have its genome completely sequenced (The C. elegans Sequencing Consortium 1998). The combination of a well annotated genome and the knowledge of the fate of every cell makes *C. elegans* a powerful system to explore and understand metazoan development as well as drug target identification and validation. Other nematode genomes are now available which makes this model uniquely suited for the study of comparative genomics and genome evolution.

We set out, using bioinformatics and the *Caenorhabditis* system, to investigate a fundamental question of genome evolution: introns. Despite their overwhelming presence in all eukaryotes, but especially in vertebrate genomes (Roy and Gilbert 2006), introns—the often over looked parts of genes, that separate exons and are spliced out before translation—remain surrounded in mystery. However, research has shown these genetic hitchhikers may be more than simply silent passengers and, in fact, may affect gene

expression and the shape of the genome. They can be costly as well; mutation in splice sites play a role in many human diseases such as Alzheimer's (Janssen 2000), cystic fibrosis (Niksic 1999), and some cancers (Tahira 2011), to name a few.

Introns are the intermittent but integral part of genes that get spliced from the pre-mRNA transcript. Although they are not represented in the resulting protein sequence, introns have played an influential role in the regulation of gene expression, functional diversity, and genome and protein evolution (Roy and Gilbert 2006). Intron presence is found across all eukaryotic genomes although they vary in frequency and size. Some protists have less than one-hundred introns in their entire genomes while vertebrates may have over one-hundred-thousand (Roy and Gilbert 2006). The average number of introns per gene in *Drosophila* is 4.7 and the average intron length is 1,482 bp while the human genome has an average of 7.7 introns per gene with an average size of 4,800 bp (Lynch 2007). Despite knowledge of their presence in genes, many questions remain about the origins, mode and tempo of intron evolution. Studying these questions may not only resolve some of the mystery surrounding introns but also tell us more about the structure and evolution of genes.

Two theories of the origin of introns prevail in the literature termed introns early and introns late. The introns early model states that introns were present and played a formative role in the earliest of genes through exon-shuffling. While retained in eukaryotes throughout evolution they were lost in prokaryotes (Doolittle 1978). The introns late theory proposes that introns were introduced only after the divergence of prokaryotes and eukaryotes (Logsdon 1998).

A first look at intron phase bias—the part of the codon in which the intron is positioned—has been used to support the introns early theory. Specifically, the majority of extant introns are found in phase 0 (between codons). The argument for the introns early theory holds that this position would have been the least disruptive to the reading frame during exon-shuffling (Long 1999). However, arguing against the introns early theory is the fact that most recently gained introns also show phase 0 bias. Perhaps phase bias reflects nucleotide sequence preference rather than an artifact of early intron evolution. In fact, the sequence MAG*GT (where M = A or C and * = the position of the intron) has been shown to be the preferred sequence of intron positioning within the genome (Qui 2004). Disruption of these splice sites can impede intron splicing (Aebi 1986) which suggests sequence bias is more likely the cause of phase preference than reading frame disruption during exon-shuffling.

Regardless of whether introns arose early or late, we know that new introns are still appearing and disappearing from genomes. There are currently two suggested mechanisms of intron loss:

- Reverse transcription-mediated intron loss (RTMIL) – According to this model, the transcript of the mature mRNA of a gene is reverse transcribed and converted with the gene from which it originated. Because the introns of mature mRNA have already been spliced out, the converted gene would suffer total intron loss (Fink 1987). This mechanism does not resonate with the observation that genes typically suffer the loss of only one or some of their introns, not all of them unless this process did not encompass the entire gene. If the mRNA was only partially spliced, partially reverse-transcribed or partially converted with the original gene this could account for individual intron loss.

Reverse-transcriptase also shows a bias towards the 3' end of the gene predicting intron loss to be greater towards the 3' end of genes. Because this is a transcription-mediate process, this mechanism would cause heritable intron loss only in genes expressed in the germline. Therefore biased germline expression would be a good indication of this mechanism's role in intron loss.

- Genomic deletion – In this simple model, introns may be precisely or imprecisely excised from the genomic sequence (Roy 2006). Imprecise deletion would not cause disruption of the reading frame as long as any intron nucleotides left behind or exon nucleotides removed are a multiple of three and do not contain a stop codon. While precise intron deletion leaves no trace behind to provide a clue for the cause of the loss, imprecise deletion would affect the nucleotide sequence surrounding the intron loss site. If genomic deletion is a leading cause of intron loss in our dataset, we expect to see a disproportionately greater number of intron loss in the untrimmed dataset—where no effort is made to eliminate ambiguous alignments—than in the trimmed dataset.

While there are only two proposed mechanisms of intron loss there are many more potential mechanisms of intron gain put forth by the literature. Seven models have been proposed to explain intron gain:

- Intron transposition – According to this model, an intron spliced from the transcript of a gene is reinserted either into a new location in the same transcript or into a different transcript entirely. The mRNA then undergoes reverse transcription followed by gene conversion (Sharp 1985). Although the process of intron transposition might seem like it relies on a series of serendipitous steps and thus unlikely to be very common, there have been findings in some studies that support this process (Baltimore 1985; Sharp

1983). For the resulting intron-gain to be a permanent new feature of the gene, the gene would have to be expressed in the germline. A bias of germline expression among the genes with recently gained introns may implicate intron transposition in the role of recent intron gain.

- Tandem genomic duplication – This model proposes that intron gain occurs when a region of a gene containing the sequence AGGT is duplicated, creating new splice sites recognized by the spliceosome. The sequence of the novel intron, surrounded by new splice sites would be a duplication of the adjacent sequence (Rogers 1989). Depending upon where the splice site is in the duplicated DNA segment, tandem genomic duplication may be either precise or imprecise. If it is precise, the coding sequence of the gene remains unaltered; if it is imprecise, the coding sequence may be altered. If BLASTing the recently gained intron sequences against their own genome reveals matches that are right next to the original intron, tandem genomic duplication might have been the mechanism that created the newly gained intron.
- Intronization – The model of intronization proposes that mutations within the exon sequence may generate new splice sites recognized by the spliceosome. The sequence between the novel splice sites would be excised as a novel intron (Irimia 2008). Alternatively, Catania and Lynch hypothesized that this process is gradual and may involve alternative splicing (2008). They propose the fortuitous splicing of a premature termination codon-containing exon may become the dominant splice variant. Over time mutation in the splice sites may allow spliceosomal recognition creating a novel intron. If intronization caused an intron gain, the resulting peptide length should be shortened. We

can look for evidence of this by comparing amino acid sequence length of the genes containing recent intron gains.

- Intron transfer – In this model, paralogs align and undergo recombination of an intron containing site with an intronless site (Hankeln 1997). An imprecise insertion event may be tolerated better in an unessential gene copy, causing the observed higher rate of intron gain among paralogs (Babenko 2004). Over time mutation and recombination may allow more precise insertion of the intron in the functional copy of the gene. Evidence of this mechanism may be found among the BLAST results if any of the BLAST hits match the sequences of introns in other paralogs. While our ortholog set does not contain any orthologs that have a paralog, it is possible that there may be signs of intron transfer among genes of the same family.
- Self-splicing type II intron – The mitochondria of many eukaryotes contain self-splicing type II introns. According to this proposed mechanism, it is possible that a self-splicing intron from a mitochondrial gene could migrate to the nucleus where it would insert itself into a new gene and undergo conversion into a spliceosomal intron (Cavalier-Smith 1991). Modern spliceosomal introns share many traits with self-spicing type II introns such as the sequences on the 5' and 3' ends. Given their similarities it is not unreasonable to guess that type II introns may be the primitive ancestors of some spliceosomal introns. BLASTing recently gained introns against their own genome might reveal if any of the recent gains matched a mitochondrial sequence. This mechanism is not relevant to this study because *Caenorhabditis* mitochondria do not contain self-splicing introns.
- Transposon insertion – This model proposes that a transposable element

(TE) inserts itself into a gene in an AG*GT sequence and transforms into an intron that is able to be recognized by the spliceosome (Purugganan 1992). The theory of intron gain by transposon insertion is neither new nor without precedent. Already there has been some plausible, although indirect evidence of TEs causing intron gain in *C. elegans* (Roy 2004). There is even conjecture that TEs may be somewhat responsible for the phenomenon known as intron drift by leaving behind a small nucleotide segment that can act as a new 5' splice site (Lynch 2007). BLASTing the known transposon sequences against the list of recently gained intron sequences should reveal any evidence of transposon activity.

- Double strand break repair – (DSBR) A recent study has suggested a potential novel mechanism of intron gain: a staggered double strand break may be repaired by the insertion of small fragments of DNA, effectively creating a new intron (Li 2009). This mechanism predicts the presence of small repeated sequences surrounding the recently gained intron.

In their paper, Li et al. compared the intron boundaries in the genomes of eighty-four isolates of *D. pulex* (2009). They were able to identify 24 cases of intron variation. An astounding 87.5% of these (21/24) were putative intron gains with only three losses. Despite using such closely related genomes, they were unable to find any evidence that might suggest any of these gains were caused by the traditional theories of intron gain. What they did find however, were short segments of repeats (5-12bp) that flanked the gained intron.

These tandem repeats suggest a new model of intron gain: double stranded break repair. Li et al. hypothesized that during the process of repairing these staggered, double

stranded breaks, a small, random segment of nucleotides is inserted between the break. Novel introns may in fact be the simple result of DNA repair.

Until now, there hasn't been a study done on intron evolution with this fine a resolution. By comparing isolates from the same species, Li et al. were able to identify a previously unseen mechanism of intron gain. It is highly likely that these new introns caused by double-stranded repair may occur in other species but the research comparing that level of homologous genomes has not been done yet.

Previous studies in *Caenorhabditis* have also turned up evidence of gains but the mode of these remain unclear due to the phylogenetic distance and the rate of silent mutation of intronic sequences (Logsdon et al. 1998). In 2004, Coghlan and Wolfe compared 12,155 orthologs in *C. elegans* and *C. briggsae* using a distantly related species *B. malayi* as an out-group. In their data set they were able to identify 41 gains in 39 genes from *C. briggsae* and 81 gains in 74 genes from *C. elegans*. However, a later study carried out by Roy and Penny disputed these numbers, saying many of those intron gains were in fact intron losses (2006). Roy and Penny used a more complete phylogenetic tree that included *C. brenneri* and *C. remanei* (fig. 1) and this allowed them to more accurately define gains and losses according to parsimony by comparing the more closely related species. For example, what Coughlan and Wolfe might call a gain in *C. elegans* because of its absence in the *C. briggsae* ortholog, might actually be an intron loss in *C. briggsae* if the intron is also found in the same position in the *C. brenneri* and *C. remanei* orthologs. In this later research, Roy and Penny found that 74% of the gains in *C. elegans*, while absent in *C. briggsae*, were present in the other two *Caenorhabditis*

Figure 1:

Phylogenetic Tree of *Caenorhabditis*

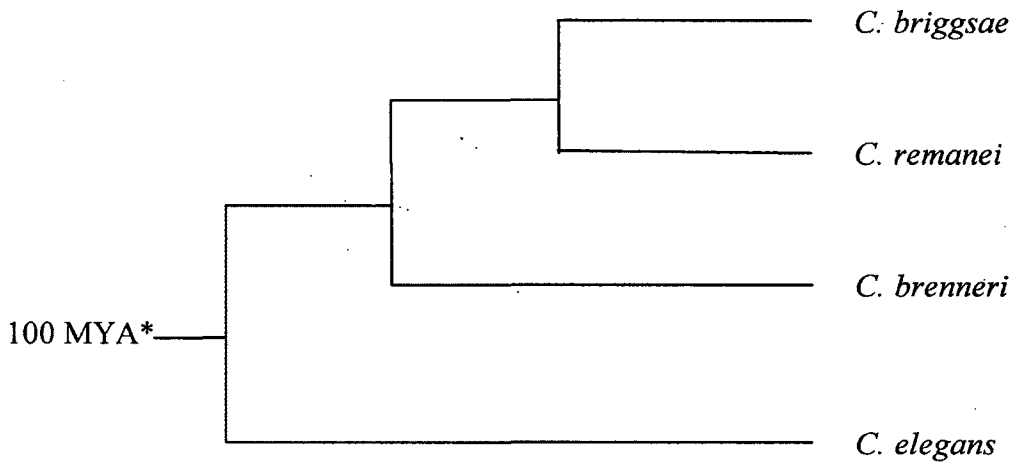


Figure1 Phylogenetic tree of *Caenorhabditis* built by comparing sequences from 18S rDNA (Fitch 1995). Branch lengths have no significance. *million years ago (Gupta 2007).

species. A more comprehensive analysis of intron evolution among multiple, closely related *Caenorhabditis* species will be highly informative for polarizing gains and losses.

Another, often overlooked, facet of intron evolution is intron sliding, also known as 'intron drift' (Stoltzfus 1997). Until the year 2000, research into intron sliding was hindered by the fast rate at which intron sequences lose their sequence identity. In fact, since so little evidence of intron sliding could be provided, it was believed to be a rare occurrence if it existed at all (Rogozin 2000). And yet, intron sliding could explain some of the diversity of intron positions in orthologs. One paper hypothesizes that alternative splicing could provide a means by which intron sliding occurs (Tarrio 2008). They propose a mechanism by which strong and weak splice site junctions may switch during a speciation event. The result would be a slight shift in the intron position to the new preferred junction. With the use of more closely related genomes, it is possible to unambiguously align intron positions within coding sequences among thousands of orthologs. This along with comparisons of intron sequences could provide clear examples of intron sliding events.

So far, previous studies have only shown relatively small amounts of intron gain and loss using only distantly related species. While several mechanisms have been presented, few have many clear examples in the literature. With the exception of the study of *D. pulex* (Li 2009), our research has been conducted on a finer resolution by comparing four closely related species yielding a greater amount of polarized information than gathered previously. There are three major advantages to the *Caenorhabditis* comparison: 1) the low level of paralogy, which allows for the prediction of many pan-orthologous introns; 2) the relatively close relationships among the species, which affords

the unambiguous alignment of most intron positions in the orthologs; and 3) all genomes compared are completely sequenced.

Using a bioinformatics reciprocal best BLAST approach, we compare intron positions in 6,257 orthologs from four different *Caenorhabditis* species (*C. elegans, C. brenneri, C. remanei* and *C. briggsae*). In this manner we can measure the rate of intron gain and loss by utilizing the known phylogenetic relationships. Because these species are closely related we have a better opportunity to not only map gain/loss of introns but also possibly identify the origins of recently gained introns as well as uncover intron sliding events. Of course this presumes, according to uniformitarianism, that the mechanisms and patterns of intron gain and loss have not changed since their ancient origins.

MATERIALS & METHODS

Starting Datasets

The input-data files, which include the Genome Feature File (GFF), the Genome Transfer Files (GTF), the contig files and the amino acid files, were downloaded from wormbase.org's FTP site. We used the WS_95 version of the C. elegans GFF and the corresponding AUGUSTUS versions of GTFs for the other three species.

Orthologs

The starting number of 6,546 orthologs was derived using reciprocal best BLAST (Phil's paper). Only orthologs without paralogs were allowed. Orthologs that had one or more species with a frame error in their sequence were removed from the set.

Intron Positions

Using the positional information from the GFF or GTFs, we constructed the CDS (coding sequence) from the contig file. Exon-exon junctions were separated with a slash to bookmark the intron position. If a gene had an alternative splicing, the splicing variant that produces the longest CDS was selected. In some cases of alternative splicing, only the UTRs are affected and the CDS remains the same length in all splicing variants. In this instance, the first splice variant was selected (Zahler 2005).

The resulting FASTA files, each containing one species' orthologous coding sequences with inserted slashes, were translated into amino acid sequences. The slashes

were converted to one of three special characters depending upon the intron phase and were placed after the amino acid in which the intron was found. Clustal does not accept numbers and so the letters O, U and Z were used to represent phase 0, 1 and 2 respectively. We chose the letters O,U and Z because they do not represent any amino acids among our data set.

The modified amino acid sequences were grouped into FASTA files by ortholog group.

Sequence Alignment

The FASTA files were run through Clustal for multiple-sequence alignment, using the default parameters (Larkin 2007). The output order was locked to maintain the same order of the inputted sequences, rather then reorder by alignment strength.

Because Clustal does not recognize the special characters and thus does not make any efforts to align them, we wrote a script to insure intron position conservation over areas with gaps.

Assigning Gain and Loss

We developed an algorithm and code that analyzes the modified amino acid alignments for gain, loss, conservation or ambiguity of intron position based on the guide tree. An intron found in the same position and phase in only one of the species but not the others was considered a gain. An absent intron in any of the species where the other three had an intron was considered a loss. If an intron was present in only *C. briggsae* and *C. remanei* it was considered a gain in those two species or a loss if absent from only those

two species. Otherwise, if there was an intron present or absent in any other two combination of species (*C. briggsae* and *C. elegans*, *C. briggsae* and *C. brenneri* or *C. brenneri* and *C. remanei*) it was considered ambiguous. An intron found only in *C. elegans* but absent in the other three species can not be called a gain because it is the out group in our dataset. Likewise an intron absent in *C. elegans* but present in the other three species can not be called a loss.

Intron variation within 5 codons was excluded in an attempt to account for the possibility of drift. We wrote a script to sort through the gain/loss results and print any variation that occurs within five codons to a separate file. Conserved intron positions and intron variation occurring outside five codons were printed separately.

Trimming Alignments

In order to examine only unambiguously aligned intron positions, we used a program developed by Phil Hatcher to walk through the alignments and cut out any ambiguously aligned regions. This was done by setting the number of gaps and mismatches allowed between walls. A wall is a position containing only conserved amino acid residues. A column may be a line containing one or more—depending on the setting—mismatched amino acid. Three trim settings were used. A "relaxed" trim was defined by setting the column to 0.1, which allows one gap in any of the four species, and the wall to 1.0, which allows no gaps. A "strict" trimming was defined when the column and wall are both set to 1.0. An "X" was used as a place holder for gaps in positions that contained an intron so as to not unnecessarily remove gained or lost introns from an otherwise conserved sequence.

To further reduce the ambiguity of the trimmed data, we checked the alignments by hand to confirm intron conservation over gaps. Because of this the number of conserved introns in both trimmed data sets increase slightly when compared to the untrimmed data.

Confirming Intron Annotation

We compared exon-exon junctions of the most recently sequenced genomes, *C. brenneri* and *C. remanei*, with cDNA data in order to identify any possible annotation errors. Exon-exon junctions of all ortholog genes in *C. brenneri* and *C. remanei* were extracted from the contig file using the coordinates in the GTF. The exon-exon junctions consisted of twenty nucleotides from the tail end of the upstream exon and the start of the downstream exon. The entire sequence was 40 nucleotides in length. These small junctions were BLASTed against the EST database downloaded from NCBI site (Boguski 1993). No cut-off e-value was given but only the first/best result per query was allowed.

Intron Gain BLASTs

The intron sequences of all the putative gains were BLASTed against their own species. The e-value cut off was 10^{-100}. The results were parsed in such a way that any hit overlapping the original intron position was removed. The results were then annotated by hand by finding the chromosome location in the GTF or in the case of *C. elegans* using wormbase.

Relative Positions of Intron Gains and Losses

Intron gains and losses from the untrimmed data were graphed by relative position in the alignment of the gene. The untrimmed data was used because the trimming program loses intron-position information that is necessary for this analysis. The position of the intron in the alignment of the CDS was divided by the length of the alignment in order to categorize introns by their relative position.

Germline Expression

A list of gene IDs of genes containing one or more gained introns was generated. The *C. elegans* ortholog of these genes was used to search the wormbase database for expression in the germline using Wormmart. The WS190 database was used and the "expression pattern" dataset was selected. The anatomy term "WBbt:0005784" was set as a filter and the file of geneIDs was uploaded.

RESULTS & DISCUSSION

Orthologs in *Caenorhabditis*

We set out to investigate intron gain and loss in *Caenorhabditis* using a comparative genomics approach. We used the completely sequenced and annotated genomes of four species of the small nematodes: *C. elegans*, *C. brenneri*, *C. briggsae* and *C. remanei*. Using a starting ortholog set of 6,546 orthologs predicted by reciprocal best BLAST (RBB), we looked for intron presence/absence.

Previous studies with few exceptions have only compared intron positions in this many orthologs, from distantly related species such as *C. elegans*, mouse and human. Due to the rapidly growing area of genome sequencing, we are now able to compare the genomes of much more closely related species. This allows a more confident assessment of intron position among the ortholog alignments. We made a further refinement of the alignments by eliminating sections with a lower percent identity as described in the methods section as well as later in the text.

Out of the starting set of 6,546 orthologs, 289 were omitted from the study due to annotation errors as described in the methods section, resulting in a final total of 6,257 orthologs. Among the *Caenorhabditis* species the number of genes range from 21,391 in *C. briggsae* to 43,238 genes in *C. brenneri* (table 1). The large number of orthologs presents an opportunity for finding trends in intron gain and loss that might otherwise be unnoticeable in a smaller data set.

Increased intron density in *Caenorhabditis* orthologs

Within the 6,257 orthologs the number of introns average 40,094 per species with *C. brenneri* having the least at 38,470 and *C. elegans* having the most at 42,561. The average number of introns per gene ranges from 6.1 to 6.8 among the ortholog set. However, the average number of introns per gene in the entire genome is 4.1 introns/gene in all four species (table 1). This is a significantly greater average of introns per gene in the ortholog set compared to the entire genome. There might be a logical explanation for the difference in average intron/gene. It could simply be indicative of an ascertainment bias in the ortholog set; perhaps orthologs, by their nature of homology, have more clearly annotated introns. On the contrary, there may be some inherent characteristic of these orthologs that predispose them to this greater intron variation.

Ortholog sequence alignments

In order to compare intron positions, we aligned the ortholog sequences using command line Clustal W2. Following this, we used a script called cutter.pl (see appendix) that removes sections of the alignments based upon input parameters. Three levels of stringency were applied. The least stringent level was simply the raw alignments as they were produced by Clustal (untrimmed). The second level of stringency eliminated sections of alignments containing more than one gap (trimmed), while the third level allowed for no gaps in the alignments (strictly trimmed).

Table 1:

Summary of Intron Numbers in *Caenorhabditis*

	Genes	Introns		Ortholog Introns		CDS Introns		After Trim Introns	
		total	per gene	total	per gene	total	per gene	Total	per gene
C. brenneri	43,238	166,811	3.9	38,470	6.1	38,404	6.1	33,659	5.4
C. briggsae	21,391	97,479	4.6	39,095	6.2	39,082	6.2	33,545	5.4
C. elegans	26,654	104,916	3.9	42,561	6.8	42,307	6.8	35,835	5.7
C. remanei	33,678	136,164	4.0	40,249	6.4	40,209	6.4	34,414	5.5

Table 1 A table showing the total numbers of introns in each *Caenorhabditis* species' genome, ortholog set, coding sequences within the ortholog set and the final number of introns in the dataset we used for comparison. The ortholog intron column contains introns in the entire gene while the compared-CDS column contains only introns found within the start and stop codon. The number of introns compared in the raw CDS differ from the number of introns in the trimmed CDS due to the removal of ambiguous regions of the CDS that may or may not contain intron positions.

Intron variation

A primary goal of this research is to identify intron gains and losses in four *Caenorhabditis* species. In order to accomplish this we placed numerical markers in the amino acid sequences to represent an intron position. The numbers correspond to the intron phase, or in other words its relative position within the codon. Phase 0 denotes the position between codons; phase 1 is between the first and second nucleotides of the codon; phase 2 is between the second and third nucleotides. After alignment of the orthologs the intron markers were assessed for presence/absence using the phylogenetic guide tree shown in figure 2.

An intron is considered conserved if it is in the same position and phase in the amino acid sequence of all four species. Gains and losses are categorized for all species with the exception of *C. elegans* which is the out group in this analysis. Given its phylogenetic relationship to the other *Caenorhabditis* species we cannot predict the direction of intron variation in *C. elegans*. However, for the other three species, a gain is any intron found in a position in the amino acid sequence of only one species but not the other three. Based on parsimony, the intron is also called a gain if it is found in the same position and phase in *C. briggsae and C. remanei* but not in the other two species. A loss is the absence of an intron in the amino acid sequence of one species where an intron was found in the other three species. A loss is also noted if *C. briggsae* and *C. remanei* lacked an intron in the same position where the other two species contained one. Any other permutation of intron presence/absence among the four species is marked as ambiguous.

We found 27,927 conserved intron positions in the most stringently aligned dataset (table 3). Almost half of the intron positions are conserved in phase 0; a quarter of the positions are conserved in phase 1 and a quarter are conserved in phase 2. The resulting ratio is 2:1:1 for phases 0, 1 and 2 respectively. This phase bias holds true also for the intron gains and losses in the most conserved data set. This observation supports the theory of introns preferring certain splice site sequences rather than being supportive of either the early or late theories of intron evolution.

In summary, the majority of intron positions are conserved and overall there is greater intron loss than intron gain. In this aspect our intron gain and loss results agree with the trends seen in previous studies (Coghlan 2004; Coulombe-Huntington 2007). However, unlike in previous studies we see a greater amount of intron gain in our results. It has been argued that there is only a minimal amount of recent gain in *Caenorhabditis* (Roy and Penny 2006). Even under the most stringent criteria, our results have a significant amount of recent intron gain. The close phylogenetic relationship among the four *Caenorhabditis* could explain the greater amount of intron flux seen in our data set. Intron positional variation that might otherwise be lost over a longer separation of species is more evident amongst genes that have had less time to diverge.

Figure 2:

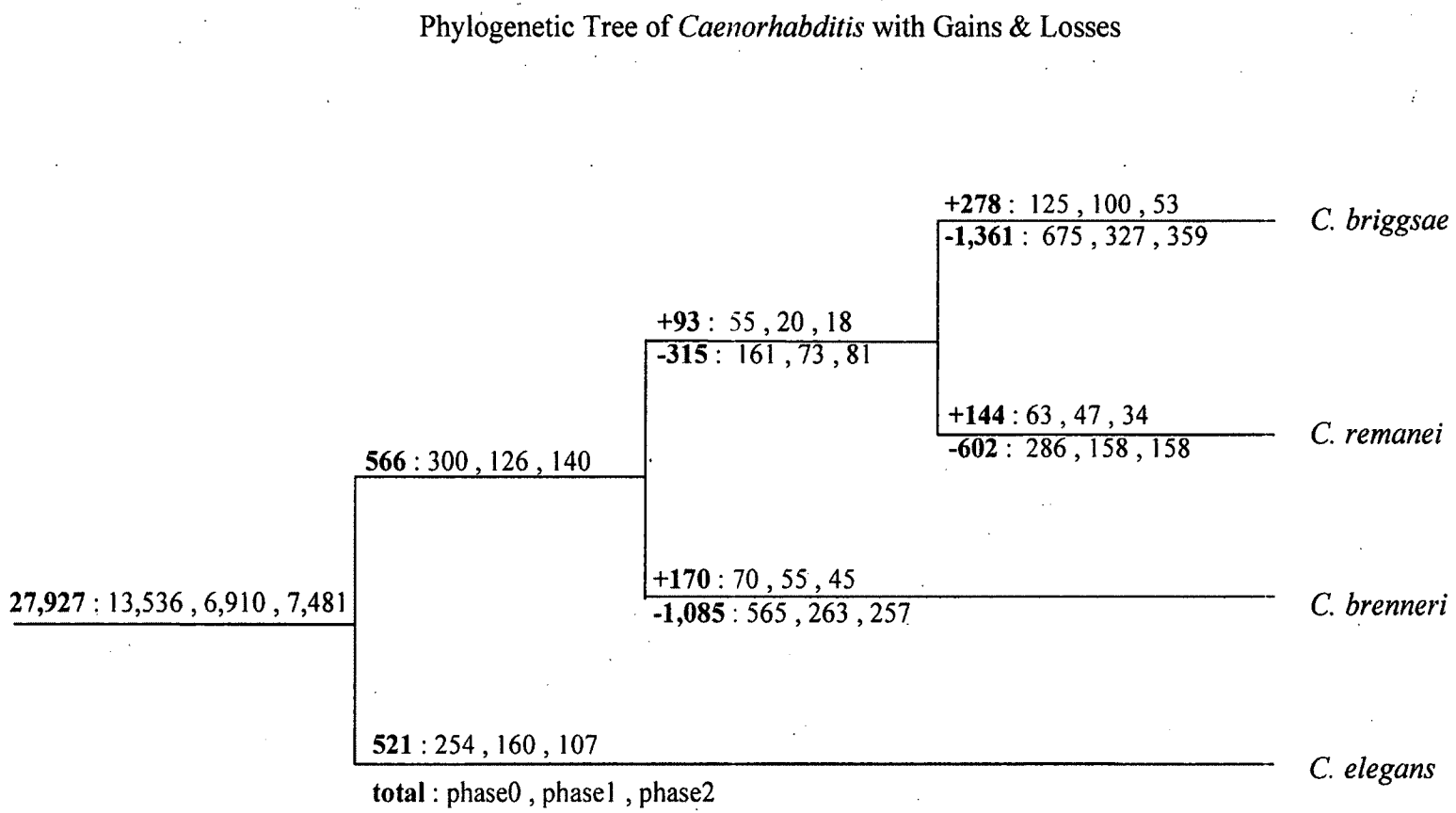


Figure2 Phylogenetic tree of *Caenorhabditis* (Fitch 1995). Branch lengths have no significance. Total gain and loss of intron positions are noted for each branch as well as parsed by phase in the order 0, 1 and 2. The gain and loss numbers are from the strictly trimmed dataset excluding intron variation within 5 codons.

Lineage specific intron variation

Another observation previously unnoticed is a lineage specific rate of gain and loss. As seen in figure 2 and table 2 & 3, *C. briggsae* has greater numbers of intron gain and loss than *C. brenneri* and nearly twice the amount of gain and loss when compared with *C. remanei*. These differences occur despite the total number of introns in the ortholog set being relatively similar among all four species.

One cause of the overall difference in intron gain and loss between species could be that some mechanism of gain and loss is more active in one species compared to the others. Another possible explanation could be population size. The wild population of these *Caenorhabditis* species is not known, but it is possible that *C. briggsae* has a smaller population than the other *Caenorhabditis* species. A theory of how population size affects intron loss and gain is described by Michael Lynch in his book The Origins of Genome Architecture. He proposes that a larger population is less tolerant of a large intron burden compared to a smaller population. Introns come at a cost because they are prone to unchecked mutation that can have deleterious effects. They also cost more energy during genome replication. In smaller populations it is more likely that a compromised individual would, by chance, reproduce despite the risks and drawbacks of a large burden of introns. In a larger population it is less likely for such an individual to sustain a genetic footprint in the population. If the *C. briggsae* population is smaller than the other species this theory could explain the greater number of intron gains and losses.

Table 2:

Intron Gain and Loss

Untrimmed Raw

phase	eleg-only	not-eleg	bren gain	bren loss	brig gain	brig loss	rem gain	rem loss	consv.	ambig	phase
0	1,640	787	1,282	1,782	1,963	2,308	1,406	1,441	12,322		
1	1,200	418	1,033	928	1,428	1,199	1,078	819	6,300		
2	878	377	763	855	1,094	1,141	785	680	7,125		
total	3,718	1,582	3,078	3,565	4,485	4,648	3,269	2,940	25,747	3,051	10

Trimmed Raw

phase	eleg-only	not-eleg	bren gain	bren loss	brig gain	brig loss	rem gain	rem loss	consv.	ambig.	phase
0	711	555	486	1,169	748	1,551	529	925	13,536		
1	525	281	447	632	693	873	468	586	6,910		
2	401	275	358	555	554	845	340	474	7,481		
total	1,637	1,111	1,291	2,356	1,995	3,269	1,337	1,985	27,927	1,516	178

Trimmed-strict Raw

phase	eleg-only	not-eleg	bren gain	bren loss	brig gain	brig loss	rem gain	rem loss	consv.	ambig.	phase
0	271	310	92	585	215	862	139	462	13,536		
1	171	130	79	278	146	427	91	245	6,910		
2	118	147	60	268	98	459	63	258	7,481		
total	560	587	231	1,131	459	1,748	293	965	27,927	616	6

Table 2 The table contains the number of gained, lost, conserved or ambiguous intron positions for each species by intron phase. The intron positions in *C. elegans* cannot be described as gained or lost positions due to the species being the out-group. The table lists the *C. elegans* intron positions as those introns found only in *C. elegans* or those introns found in the other three species but not *C. elegans*. Some positions were considered ambiguous when a conclusion of gain or loss could not be reached based upon the phylogenetic tree. The phase column describes intron positions that vary only by codon phase. The amount of conserved positions as well as the phase column differ from the untrimmed data and the trimmed data because the trimmed alignments were reviewed by hand to assure intron position conservation surrounding gaps.

Excluding intron variation within 5 codons

One unexpected pattern became clear upon a more detailed look at the aligned sequences: the presence of what may be called intron drift or sliding. The phenomenon known as drift occurs when an intron position is found only a few codons away from its corresponding intron position in the other species. Although previously thought to be very rare, our results show an unprecedented amount of potential intron drift. We defined potential drift to occur when intron positions are within 5 codons of each other.

In order to accommodate the possibility of drift affecting our results, we put the intron gain and loss numbers through yet another filter that excluded intron variation occurring within five codons of each other. The assigned five codon distance is an arbitrary number chosen by convention; introns may drift further or less than five codons and thus the actual number of drift might be lower or higher than our results. Furthermore, some or all of these introns may be separate gain and loss events and not drift at all.

In the trimmed data, sections of ambiguous alignments are removed which may artificially cause two intron positions to fall within 5 codons on each other. Therefore, the numbers of potential drift identified by proximity may be slightly inflated in the trimmed and strictly trimmed data sets.

In the untrimmed data set there were 1,391 cases of potential intron drift in *C. elegans*, 1,555 cases in *C. brenneri*, 2,432 cases in *C. briggsae*, 1,759 cases in *C. remanei* and 764 cases classified as ambiguous (table 3). Overall there were 7,901 cases of suspected intron drift in the untrimmed data set.

Are these introns suspected of drifting authentic case of drift? These potential drifted introns could be the result of two consecutive events: a loss and a subsequent gain (or vice versa), but there is the chance they could be the same intron only slid slightly up or down stream from its original position. To avoid ambiguity in our gain and loss results, any intron variation occurring within five codons was filtered from the results. The filtered gains and losses differ from the unfiltered or "raw" data only in volume. The pattern of gain and loss remained the same; the majority of the intron positions were conserved and there was greater intron loss than gain. The gain and loss pattern reflects the raw data which implies that whether or not these introns are drifters does not affect any conclusions we make about our gain and loss numbers.

Table 3:

Intron Gain and Loss Excluding Variation within 5 Codons

Untrimmed

phase	eleg-only	not-eleg	bren gain	bren loss	brig gain	brig loss	rem gain	rem loss	consv.	ambig	phase
0	1,201	633	830	1,576	1,304	1,896	891	1,164	12,322		
1	868	301	659	797	901	958	679	653	6,300		
2	617	289	483	743	698	944	508	555	7,125		
total	2,686	1,223	1,972	3,116	2,903	3,798	2,078	2,372	25,747	2,287	10
var	1,391		1,555		2,432		1,759		0	764	0

Trimmed

phase	eleg-only	not-eleg	bren gain	bren loss	brig gain	brig loss	rem gain	rem loss	consv.	ambig.	phase
0	590	502	380	1,114	581	1,437	410	860	13,536		
1	432	253	337	582	523	771	346	519	6,910		
2	318	246	259	510	399	753	238	417	7,481		
total	1,340	1,001	976	2,206	1,503	2,961	994	1,796	27,927	1,300	178
var	407		465		800		532		0	216	0

Trimmed-strict

phase	eleg-only	not-eleg	bren gain	bren loss	brig gain	brig loss	rem gain	rem loss	consv.	ambig.	phase
0	254	300	70	565	180	836	118	447	13,536		
1	160	126	55	263	120	400	67	231	6,910		
2	107	140	45	257	71	440	52	239	7,481		
total	521	566	170	1,085	371	1,676	237	917	27,927	584	6
var	60		107		160		104		0	32	0

Table 3 A table of intron positions gained, lost, conserved or ambiguous, excluding variation in intron positions when found within five codons of each other. The row titled "var" is the total number of varied intron positions found within 5 codons of each other for that species.

Intron positions confirmed with EST data

One concern about the alignments is that the prevalence of apparent intron drift could be a sign of annotation error in the species whose genomes are newly sequenced and less well annotated. *C. elegans* has a well-established, and thoroughly sequenced and annotated genome and *C. briggsae* is not far behind. It is unlikely that the intron positions in these lineages suffer from annotation error. The genomes of *C. brenneri* and *C. remanei* on the other hand, were relatively recently sequenced and are not yet quite as thoroughly or reliably annotated. If annotation is a contributing factor to the relatively large number of suspected cases of intron drift, we would expect to see a lower amount of suspected drift in *C. elegans*. And in fact, in the untrimmed data the other three species of *Caenorhabditis* have an average of 1,915 intron positions that varied within five codons compared to *C. elegans* which has 1,391 varied positions (table 3). This could also imply that some of the apparent drift in the other, more recently sequenced species is due to annotation error. In order to establish a level of confidence in our results, we searched for the exon-exon junctions of the orthologs from the two more recently sequenced species—*C. brenneri* and *C. remanei*—in an EST database.

We constructed a list of exon-exon boundaries for all of the orthologs in *C. brenneri* and *C. remanei* and BLASTed them each against their own EST database. The length of the exon-exon boundaries is described in the methods section. If the annotation of the intron position is accurate, the expected result should be a full-length match to an EST with 100% identity or no match found at all if there simply was no EST data for that particular junction. An imperfect hit would suggest a miss-annotated intron position.

The majority (84% in *C. brenneri*, 83% in *C. remanei*) of the exon-exon junctions BLASTed against the EST databases did not find a match anywhere in the database (table 4). This is not unexpected because the EST database is not complete and therefore may not contain the matching EST of the exon-exon junctions queried. Conclusions of the accuracy of the exon-exon boundaries that found no match cannot be made. However, 5,534 exon-exon junctions in *C. brenneri* and 5,552 junctions in *C. remanei* found full length, perfect matches in the EST database confirming their annotation. Only a small minority of junctions (1% in *C. brenneri* and 3% in *C. remanei*) found either perfect but not full length or imperfect matches suggesting only a small number of introns out of the ones with ESTs in the EST database are missannotated in these lineages.

The subset of confirmed, correctly annotated intron positions were then used to examine intron gain and loss. The pattern of gain and loss in this subset was not found to be qualitatively different from the pattern of gain and loss in the overall ortholog set. Although a small portion of intron positions may be inaccurately annotated, their presence does not appear to affect the overall pattern of intron variation. It could suggest however, an explanation for some, but certainly not all, of the apparent intron drift.

Table 4:

Exon-exon Junctions Found in EST Database

	C. brenneri	*C. remanei*
100% identity & length	5,534	5,552
100% identity, 70-98% length	255	399
93-98% identity	421	954
not found	33,394	34,550

Table 4 Total number of hits from BLASTing *C. brenneri* and *C. remanei* exon-exon junctions against their respective EST databases.

Although the possible annotation errors may explain part of the intron variation that is found within the five codons, it does not explain all of it. To confirm the annotation of at least a portion of the introns suspected of drift, we searched for the putative drifted introns among the perfect exon-exon boundary EST BLAST results and we indeed found some. We confirmed the annotation of 75 out 1,132 potential cases of drift in *C. brenneri* and 56 out of 938 potential cases of drift in *C. remanei*. At the very least this subset of the introns suspected of drift, we can confidently say, is not due to annotation errors. If further research into intron drift is to be undertaken it would need to be this subset of suspected drift that should be examined.

Do suspected drifted introns share sequence identity?

Ideally, aligning the sequence of the putative drifted intron with its most closely related conserved intron counterpart, would help answer the question about the existence of intron drift. The difficulty lies in the fast-rate of mutation of intron sequences. Even between two conserved introns belonging to the most closely related species, *C. briggsae* and *C. remanei,* the alignment is poor. If a more closely related model is used—for example, different strains of the same species—the percent identity between the intron sequences might be higher and therefore it would be easier to find evidence (or lack thereof) in the intron sequence alignment of drift. The catch-22 is that the similarity of the genomes from different strains usually results in highly conserved intron positions with few cases of putative drift.

Nevertheless, we chose to align several introns by hand as anecdotal evidence for a qualitative answer to whether or not these introns could be cases of drift. The sequence

of a drifted intron found in *C. remanei* was aligned with its conserved counterpart in *C. briggsae*. For comparison the *C. briggsae* intron sequence was also aligned with another conserved intron in *C. brenneri*. Both alignments showed roughly the same amount of identity (fig 3). Despite the rapid rate of sequence variation among the introns, the putative drifted intron in *C. remanei* showed a fairly strong alignment with the conserved intron in *C. brenneri*. The alignment score was 68.09 for both alignments. It appears that at least in our example, it is very likely this is the same intron separated by one nucleotides rather than a separate loss and gain event.

Thus far, verification of intron drift on a larger scale has failed to be realized but there are at least plausible means by which drift may occur. One proposed mechanism behind intron drift concerns alternative splicing and strong and weak intron splice sites. The strong and weak intron splice sites could switch causing the intron to slide away from its original position. Although our data does not focus on the nucleotide sequences and splice sites surrounding introns, previous studies have shown the role of strong and weak splice sites in alternative splicing in *C. elegans* (Tarrio 2008). Therefore, this proposed mechanism has a valid basis in *Caenorhabditis*. Furthermore, in the untrimmed data set there were 9 conserved intron positions that were five or less codons from a novel intron in one of the four species. While seemingly an exon of such short sequence may seem unlikely there are documented cases of micro-exons in *C. elegans* (Volfovsky 2003). These rare events may be precursors of intron drift as one intron becomes the more dominantly spliced variant.

In his book The Origins of Genome Architecture, Lynch proposes another possible mechanism of intron drift (2007). He notes that Tc1 transposable elements may

alter existing introns upon their insertion and excision. When Tc1 is excised from the intron it is imprecise, leaving behind a small TGTA insert which may act as a new 5' splice site for the intron (Carr 1994). In this way it could cause slight drift of the intron-exon junction.

Figure 3:

Example of aligned introns suspected of drift

Intron position in amino acid alignment

```
C. brenneri   ...KKKDLESORREEK-HDLLNKRREQERELHGLQRKRAIIQ
C. briggsae   ...KKKDLESORREEK-HDLLNQRREHEKELLGLQRKKALIQ
C. elegans    ...KKKDFESORKEEK-HDLLNKRREQEKELKSLQRKKALIQ
C. remanei    ...KRRTWN-2RDAKRNNELLNKRREQEKELQGLQRKKAMIQ
                 *::  :  *  :: ::***:***:*:** .****:*:**
```

Alignment of intron sequences

Possible drift	*C. remanei* GTTGGAACATCAAATTTCAATA-AATTTAAGCTAACAAAAATTATTTCAG *C. briggsae* GTAAGAAC--TGAATTCATTTAGAACTAAAAATGAAACATATT-TTTCAG ** **** **** ** ** * ** * * * * *** ******
Conserved	*C. brenneri* GTAAGAATCCTTTCAAATTTCCCAACAATTTTCAA--AATTAT-TTTTTCAG *C. briggsae* GTAAGAACT-----GAATTCATTTAGAACTAAAAATGAAACATATTTTTCAG ******* **** * ** * ** ** ** ********

Figure 3 An example showing a potential drift event in the amino acid alignment and the drifted intron's corresponding sequence aligned with the closest related conserved intron. An alignment of two of the conserved introns is shown for a qualitative comparison. Intron position and phase are noted by a number in the amino acid sequence alignment. The intron in *C. remanei* is separated from the conserved position of the other species by one nucleotide. Intron sequence alignments of the potential drifter and the most closely related species *C. briggsae* are compared to an alignment of two of the conserved intron sequences, *C. brenneri* and *C. briggsae.*

Where do introns come from?

In the study of intron evolution, an important, unanswered question remains: what are the origins of spliceosomal introns? No doubt there is more than one answer to this question and there are many theories, some with more evidence than others (Roy and Gilbert 2006). We used our database of putative recent intron gains to help address this question. Because we are using four closely related species, the introns we are calling gains are only the very most recently gained introns which gives a better chance of detecting their origins. We took the intron sequences of every putative recent gain in each of the four species and used megaBLAST to search for matches in their home genome, assuming their origins are within the home genome. Although without a further out group we cannot call the introns found only in *C. elegans* true intron gains, we will treat them as prospective intron gains for this analysis. We filtered the results to include only hits that did not overlap the original intron position and exceeded the e-value cut off of 10e-100. These results symbolize only the very most recently gained introns out of a group of recently gained introns.

The results show that introns tend to either have only one match in the genome (excluding its own match) or five or more matches(table 5a). Very few introns had only two or three matches. This might indicate that introns are gained either by mechanisms that create only one new intron such as double strand break repair, or new introns arise via promiscuous elements such as transposons that cause many copies of the same sequence.

We annotated the highest scoring hit for roughly ten introns per phase per species by hand. The results are categorized in table 5b. The limitations of the three more

recently sequenced and annotated species is such that we could only discover whether or not the hit was found in an intron or an exon. If the BLAST hit matched a part in the genome that was between genes it cannot be inferred what the intron is matching, if it is matching anything of significance. In *C. elegans,* an extra category can be included because *C. elegans* has known transposable elements annotated in its genome.

The majority of hits were found to be other introns either in the same gene as the query intron or a different gene. This could be evidence of transposon activity that may create many copies through the genome. It may also be evidence of intron transfer—although our ortholog set does not include genes with paralogs, it is possible these hits are introns in other genes within the same gene family. In the case of *C. elegans* 8 introns matched a transposon suggesting that at least some of the most recent introns may come from TEs. Some of the other species' undefined intron matches may also be matching yet-to-be-annotated TEs in their own genomes. Given the rapid mutation of intron sequences, it is not insignificant that these highly scored matches are being found.

Table 5a:

BLAST results of intron gains against the genome

Num of Hits	***C. elegans*** Phase 0	***C. brenneri*** Phase 0	***C. briggsae*** Phase 0	***C. remanei*** Phase 0
One	8	14	14	14
Two	2	9	6	2
Three	0	1	2	1
Four	3	0	3	0
Five	20	0	0	2
>Five	33	5	16	13
Total	66	29	41	32
	Phase 1	Phase 1	Phase 1	Phase 1
One	3	17	23	7
Two	1	3	5	1
Three	0	5	1	1
Four	1	1	3	2
Five	2	1	0	0
>Five	15	7	12	6
Total	22	34	44	17
	Phase 2	Phase 2	Phase 2	Phase 2
One	3	13	10	5
Two	1	5	5	2
Three	1	0	2	1
Four	0	0	1	1
Fire	2	1	0	0
>Five	14	7	13	8
Total	21	26	31	17

Table 5b:

Annotated BLAST results

	Undef.	Intron	Exon	Near org. pos.	Near other gene	Transposon
C. brenneri	14	18	4	5	2	n/a
C. briggsae	7	15	1	6	2	n/a
C. elegans	14	39	2	9	3	8
C. remanei	20	18	9	4	3	n/a

Table 5a & 5b Table 5a describes the number of introns per phase per species and how many unique BLAST hits (minus the top hit which is assumed to be the intron's self in the genome) they returned. Table 5b categorizes the top BLAST hits by annotation in the GTF. *C. elegans* is the only genome that has transposons annotated in the genome.

Mechanisms of intron gain and loss

The question is not only where do introns come from but how did they get there? Although several mechanisms have been proposed, there is still little substantiation to say with any confidence how introns are both lost and gained in the genome. We have used our data to look for clues that might support any of the traditional theories of gain and loss mechanisms.

Loss - Currently there are two popular theories of the mechanisms by which introns are lost: Reverse Transcription Mediated Loss and Genomic Deletion.

Reverse transcription mediated loss - Reverse transcription mediated intron loss relies on the enzyme reverse transcriptase which is primed on the poly-A tail of mRNA. It then transcribes 3' to 5' along the mRNA, however sometimes it may not always reach the 5' end of the mRNA (Szak 2002). If RTML occurs, this characteristic of reverse transcriptase would predict a biased distribution of intron loss towards the 3' end of the gene. We searched for any evidence of this in our data set by looking at the frequency of the relative positions of intron loss in the gene. No apparent bias was observed (fig 4). The frequency of the relative positions of lost introns was evenly distributed throughout the gene. Our results in this incidence coincide with a similar study done by Cho et al. (2004) in which they found no 3' bias of intron loss positions. We also made the same examination of the relative positions of intron gains and also found no discrepancy for either end of the gene (fig 5).

In contrast, Sakurai et al. graphed the distribution of introns in intron-poor genes and found many species including *C. elegans* had a definite lack of introns at the 5' end

(2002). Thus we cannot rule out the possibility of RTML playing a role in *Caenorhabditis* intron loss.

Figure 4:

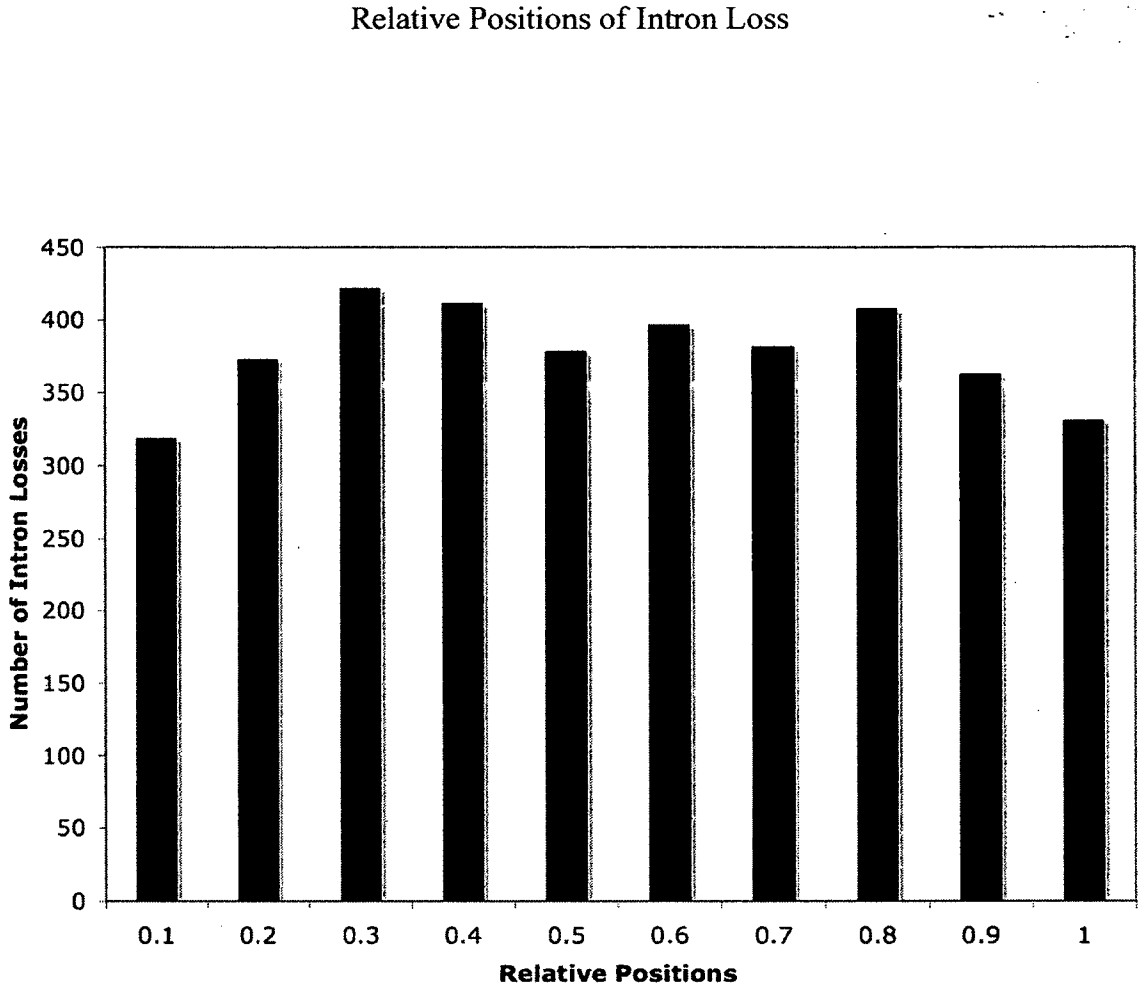


Figure 4 Graph showing the frequency of intron loss as it occurs relative to its position in the gene on a scale of 1.

Figure 5:

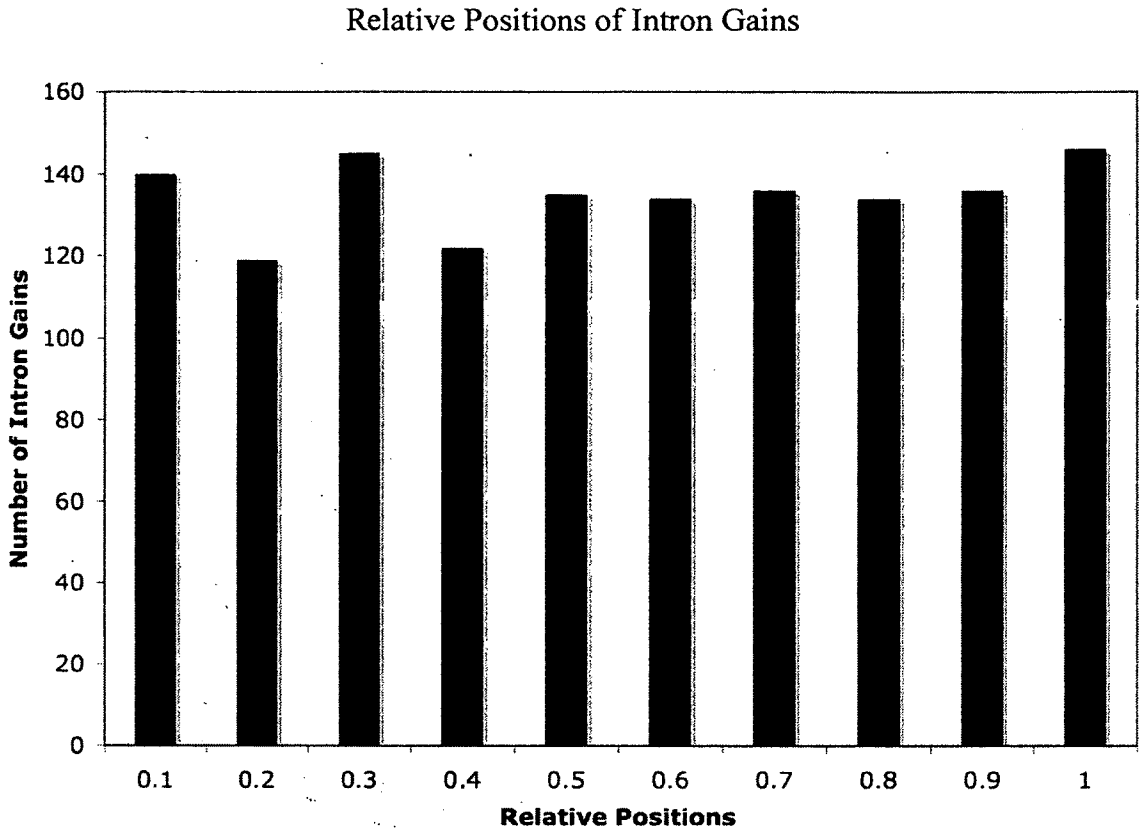


Figure 5 Graph showing the frequency of intron gain as it occurs relative to its position in the gene on a scale of 1.

Genomic deletion - Genomic deletion may be either exact in which case no trace of this mechanism would be evident or it may be sloppy. If it is the latter, we would expect to find part of the intron sequence incorporated into the reading frame resulting in a messy alignment surrounding the intron position. We can compare our untrimmed and trimmed datasets for any suggestion that genomic deletion may be a source of intron loss. In the strictly-trimmed data any sections of ambiguity were removed from the alignment. This would also remove any positions in which introns were lost due to an imprecise genomic deletion event. Therefore, we would expect to see a higher amount of intron loss in the untrimmed data compared to the trimmed in relation with the difference in gains and conserved positions.

In the comparison of the total amount of intron gain, loss and conserved numbers in the trimmed and untrimmed datasets, there is no apparent bias towards intron loss in the untrimmed data. In fact, there is nearly ten times the amount of intron gain in the untrimmed as there is in the strictly-trimmed data for the three species and only roughly three times as much loss (table 2 and 3). This does not eliminate genomic deletion as a viable mechanism of intron loss, especially because it does not account for genomic deletion that is precise. However, in our data we can see no obvious pattern that might suggest genomic deletion is a factor in intron loss in *Caenorhabditis.* The difference in gain between trimmed and untrimmed data, however may be relevant in intronization discussed later.

Although our data does not immediately reveal genomic deletion as a leading mechanism of intron loss, it may not be eliminated as a potential source of intron loss. Some convincing examples of introns being lost potentially by genomic deletion have

been found in both *Drosophila* as well as *Caenorhabditis* in previous studies (Llopart 2002, Robertson 1998).

Gains - There are seven possible mechanisms hypothesized to be responsible for intron gain. We have found some results that support a few of these mechanisms although further research is still needed. No doubt more than one mechanism may be responsible for intron gain and like intron loss, some of the mechanisms may be species-specific (Roy 2005). We have used our intron gain database to address each hypothesized mechanism.

Intron transposition - A new intron may be inserted into the gene by first being extracted from another place in the transcript or from a different transcript and then reverse transcribed into the new location. This transcript with the newly-inserted intron would then have to undergo gene conversation with the original gene for there to be a permanent effect. Although we cannot look for direct evidence of this process, in order for intron transposition to be a viable explanation for intron gain, the gain has to occur in the germline so that the new position is sustained in the genome. This is also true for RTML. If this was the case, genes containing intron gains should more likely be expressed in the germline than a random selection of genes. The *C. elegans* equivalent of genes that contained gains in the other species were searched against the *C. elegans* gene expression database on wormbase using wormmart as described in the methods section. In comparison all of the orthologous *C. elegans* genes were also searched for in the same fashion. This assumes similar expression patterns among the four species. Genes containing intron loss were also examined for germline expression.

Both sets of genes containing either intron gains or losses show a higher percentage of germline expression compared to the overall ortholog set (table 6). In other words, genes expressed in the germline are more likely to lose or gain introns than other genes. This could be evidence of both intron transposition as well as RTML as both need to occur in the germline to be viable mechanisms of intron variation. Torriani et al. in their detailed examination of three closely related species of fungi, also found indirect evidence of intron transposition (2011). They found ten recent intron gains among related gene families that had a high amount of sequence similarity and identified two possible donor loci.

Table 6:

Germline Expression of Genes with Intron Gain or Loss

Gains

	#genes	#grmline exp	%gains exp in grmline
Orthologs	6,546	39	0.6%
C. brenneri	158	4	2.5%
C. briggsae	235	9	3.8%
C. elegans	427	8	1.9%
C. remanei	130	7	5.4%

Losses

	#genes	#grmline exp	%gains exp in grmline
Orthologs	6,546	39	0.6%
C. brenneri	921	23	2.5%
C. briggsae	1,136	13	1.1%
C. elegans	496	10	2.0%
C. remanei	530	5	0.9%

Table 6 Table showing the number of genes containing intron gain or loss that were found to be expressed in the germline. The percentage of expressed genes compared to total number of genes contain gains or loss is also shown.

Tandem genomic duplication - The process of tandem genomic duplication requires an intron-containing region of a gene to be copied and inserted next to the original sequence. It is the only mechanism to be recreated *in vivo* and shown that it has the potential to create a new intron (Hellsten 2011). This mechanism predicts that recently gained introns, BLASTed against their own genome would find matching sequences directly next to their original position. Like intronization, imprecise tandem genomic duplication also causes the alignment of the amino acids surrounding the new intron to be sloppy and may also contribute to the comparability greater amount of gain found in our untrimmed data.

Among our BLAST results, an average of 12.4% of the annotated BLAST returns found matching sequences next to or near their original position (table 5b). Although the percentage is small, the results show that tandem genomic duplication holds promise as a mechanism of intron gain in *Caenorhabditis*.

Intronization - Intron gain via intronization involves the mutation of nucleotides within an exon that creates new splice sites and a new intron. This would mean the loss of a good portion of coding sequence and would result in a shortened amino acid sequence. If we compare the average coding sequence length of genes that contain intron gains between the four species it might be possible to find evidence of intronization. All four *Caenorhabditis* species have roughly the same average length coding sequences among their genes that contained recently gained introns. Only *C. brenneri* has a very slightly shorter average coding sequence length and *C. elegans* has a slightly longer average length. This could be due to many factors, one of which might be intronization.

Intronization would also leave another clue behind in the ortholog alignments. If an intron had been gained via intronization, the alignment surrounding the new intron position would be messy. A sign of this is the much greater incidence of gain in the untrimmed data set compared with loss. It could be that the over-abundance of gained introns found in messy parts of the alignments could be caused by intronization.

Intron transfer - Intron transfer is hypothesized to involve the recombination of an intron-containing portion and intronless portion of paralogs. While our ortholog set contains only pan-orthologs—orthologs that do not have any paralogs—we still may be able to find evidence of intron transfer in our BLAST results. An average of 44% of the annotated BLAST hits were found to match other introns (table 5b). Some of these other introns may belong to genes from the same family as the gene containing the putative gain and in that case these introns could be relics of intron transfer. While there has not been any evidence in nematodes of intron transfer, there has been evidence of increased rates of intron gain among paralogs in several different species (Babenko 2004).

Transposon insertion - Transposable elements are pieces of DNA that can cut or copy and paste themselves into others parts of the genome. If a TE were to jump into a gene between an AG*GT sequence, it could be treated as an intron without disrupting the reading frame. The *C. elegans* genome is well annotated which allows us to look for transposon sequences among our recent, putative intron gains. BLASTing the sequences of *C. elegans*-only introns back against the *C. elegans* genome produced 8 hits that were identified as TEs (table 5b). To examine this further we BLASTed all TE sequences extracted from the WS_95 GFF against a list of the *C. elegans*-unique intron sequences and found several very good matches.

While 81 of the TEs did not find a significant match in our list of introns, 16 TEs found one hit and 2 TEs found multiple hits all with an e-value no larger than 10^{-19}. Although the majority of these TEs belonged to the mariner family, there are also 1 Tc2-related TE, 2 Tc4s and 2 Tc3s. Eleven TEs match the same intron and while 8 of these are mariners and probably share a similar sequence, 1 is a Tc3, 1 is a Tc4 and the last match is a Tc2-related TE.

This may suggest at least a partial role of TE in the production of novel introns. Because the majority of TEs that found matches belong to the mariner family it could be an indication that these types of TEs are more likely to cause intron gain. It will be exciting to see the results of a similar BLASTs done on the other *Caenorhabditis* species as their genomes become more thoroughly annotated.

Double strand break repair - A recent paper (Li 2009) has found cases of newly gained introns in the water flea *Daphnia pulex* that appear to be the result of DSBR. This mechanism involves a small segment of random nucleotides being used like a patch to repair a double stranded, jagged break. This mechanism is detectable by the presence of small repeated segments of DNA on either side of the intron, which is what Li et al. found in the two *Daphnia* lineages.

To see if our dataset shows any cases of intron gain that could be the result of DSBR, we examined ten most recent putative intron gains from each species for those small repeats. The introns with the strongest scoring BLAST return when BLASTed against its own genome and that did not overlap the original position were considered to be the most recent intron gains. The high mutation rate of intron sequences means that an intron whose sequence is more conserved has been in existence for less time than an

intron with a more divergent sequence. The exon-intron junctions of the start and end of each recent intron gain were compared for any noticeable pattern of repetition that might indicate DSBR, but none was found. Due to the divergent nature of intron sequences, this could simply mean that even more closely related organisms (such as different strains of *Caenorhabditis*) would be required to find evidence of DSBR as a cause for intron gain.

During DSBR, mitochondrial DNA appears to be the preferred filler used in non-homologous end joining (NHEJ) in primates (Hazkani-Covo 2008). In our data however, none of our queried introns found matches in the *Caenorhabditis* mitochondrial genome, further disputing the role of DSBR in *Caenorhabditis* intron gain. On the other hand, mDNA may simply not be the preferred filler sequence of NHEJ in *Caenorhabditis*. A closer examination of intron variation among *Caenorhabditis* strains could help illuminate this mechanism.

Conclusion

The goals of this research were to identify intron gain and loss in four *Caenorhabditis* species using bioinformatics. By comparing the intron positions in orthologs, we hoped to observe patterns of intron variation that might help illuminate some of the remaining mysteries of intron origins and mechanisms of gain and loss. While the results show the majority of intron positions are conserved, there is a higher rate of intron flow than anticipated. There is also a surprising amount of putative intron drift.

These cases of gain and loss allowed us to look for any evidence that might support or refute the current, proposed mechanisms of gain and loss. There appeared to be a germline bias which supports the transcription based mechanisms of intron gain and

loss. There also is a significantly greater amount of intron gain in the untrimmed data vs. the trimmed data when compared to the loss and conserved numbers which supports the mechanisms of intron gain that cause changes in the surrounding amino acid sequence. We also found evidence of intron gain via transposon insertion.

Many mechanisms of intron gain and loss have been put forth and a few have indirect evidence of their role in intron flux. There is no reason why more than one mechanism may be responsible for intron gain and loss in the genome. Furthermore, there is no reason to believe all species gain and loss introns by the same method. It is possible that the methods of intron gain and loss differ not only by species but also over time, and that what patterns we see in modern genomes is not a reflection of the past. And so the question of intron origin remains unanswered, however we now have access to more data and a greater ability to compare that data in search of where introns came from and how they got there.

LITERATURE CITED

Aebi M, Hornig H, Padgett RA, Reiser J, Weissmann C (1986) Sequence requirements for splicing of higher eukaryotic nuclear pre-mRNA. Cell 47:555-565 doi:10.1016/0092-8674(86)90620-3

Babenko VN, Rogozin IB, Mekhedov SL, Koonin EV (2004) Prevalence of intron gain over intron loss in the evolution of paralogous gene families. Nucleic Acids Res 32(12):3724-3733 doi: 10.1093/nar/gkh686

Baltimore D (1985) Retroviruses and retrotransposons: the role of reverse transcription in shaping the eukaryotic genome. Cell 40(3):481-482 doi: 10.1016/0092-8674(85)90190-4

Boguski MS, Lowe TM, Tolstoshev CM (1993) dbEST—database for "expressed sequence tags". Nat Genet 4(4):332-333 doi: 10.1038/ng0893-332

Carr B, Anderson P (1994) Imprecise excision of the *Caenorhabditis elegans* transposon Tc1 creates functional 5' splice cites. Mol Cell Biol. 14(5):3426-3433 doi: 10.1128/ MCB.14.5.3426

Catania F, Lynch M (2008) Where do introns come from? PLoS Biol 6(11):e283 doi: 10.1371/journal.pbio.0060283

Cavalier-Smith T (1991) Intron phylogeny: a new hypothesis. Trends Genet. 7(5):145-148 doi: 10.1016/0168-9525(91)90377-3

Cho A, Jin S, Cohen A, Ellis RE (2004) A phylogeny of *Caenorhabditis* reveals frequent loss of introns during nematode evolution. Genome Res. 14:1207-1220 doi: 10.1101/gr.2639304

Cogland A, Wolfe KH (2004) Origins of recently gained introns in *Caenorhabditis*. PNAS 101(31):11362-11367 doi: 10.1073/pnas.0308192101

Collins L, Penny D (2005) Complex spliceosomal organization ancestral to extant eukaryotes. Mol Biol Evol 22(4):1053-1066 doi: 10.1093/molbev/msi091

Coulombe-Huntington J, Majewski J (2007) Intron Loss and Gain in *Drosophila*. Mol. Biol. Evol 24(12):2842-2850 doi: 10.1093/molbev/msm235

Doolittle RF (1978) Genes in pieces: were they ever together? Nature 272:581-582 doi: 10.1038/272581a0

Fink GR (1987) Pseudogenes in Yeast? Cell 49:5-6 doi: 10.1016/0092-8674(87)90746-X

Fitch DH, Bugaj-Gaweda B, Emmons SW (1995) 18S ribosomal RNA gene phylogeny for some *Rhabditidae* related to *Caenorhabditis.* Mol Biol Evol 12(2):346-358

Gupta BP, Johnsen R, Chen N (2007) Genomics and biology of the nematode *Caenorhabditis briggsae*. WormBook, ed. http://www.wormbook.org/chapters/www_genomesCbriggsae/genomesCbriggsae.html Accessed 02 October 2012

Hankeln T, Friedl H, Ebersberger I, Martin J, Schmidt ER (1997) A variable intron distribution in globin genes of *Chironomus*: evidence for recent intron gain. Gene 205:151-160 doi: 10.1016/S0378-1119(97)00518-0

Hazkani-Covo E, Covo S (2008) *Numt*-Mediated double-strand break repair mitigates deletions during primate genome evolution. PLoS Genet. 4(10):e1000237 doi: 10.1371/journal.pgen.1000237

Hellsten U, Aspden JL, Rio DC, Rokhsar DS (2011) A segmental genomic duplication generates a functional intron. Nat Commun 2:454 doi: 10.1038/ncomms1461

Irimia M, Rukov JL, Penny D, Vinther J, Garcia-Fernandez J, Roy SW (2008) Origin of introns by 'intronization' of exonic sequences. Trends Genet 24(8)378-381 doi: 10.1016/j.tig.2008.05.007

Janssen JC, Hall M, Fox NC, Harvey RJ, Beck J, Dickinson A, Campbell T, Collinge J, Lantos PL, Cipolotti L, Stevens JM, Rossor MN (2000) Alzheimer's disease due to an intronic presenilin-1 (*PSEN1* intron 4) mutation. Brain 123:894-907 doi: 10.1093/brain/123.5.894

Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG (2007) Clustal W and Clustal X version 2.0. Bioinformatics 23:2947-2948 doi: 10.1093/bioinformatics/btm404

Li W, Tucker AE, Sung W, Thomas WK, Lynch M (2009) Extensive recent intron gains in *Daphnia* populations. Science 326:1260-1262 doi: 10.1126/science.1179302

Llopart A, Comeron JM, Brunet FG, Lachaise D, Long M (2002) Intron presence-absence polymorphism in *Drosophila* driven by positive Darwinian selection. PNAS 99:8121-8126 doi:10.1073/pnas.122570299

Logsdon JM (1998) The recent origins of spliceosomal introns revisited. Curr Opin Genet Dev 8:637-648 doi: 10.1016/S0959-437X(98)80031-2

Logsdon JM, Stoltzfus A, Doolittle WF (1998) Molecular evolution: Recent cases of spliceosomal intron gain? Current Biol 8(16)R560-R563 doi: 10.1016/S0960-9822(07)00361-2

Long M, de Souza SJ, Gilbert W (1995) Evolution of the intron-exon structure of eukaryotic genes. Curr Opin Genet & Dev 5:774-778 doi: 10.1016/0959-437X(95)80010-3

Long M, Deutsch M (1999) Association of intron phases with conservation at splice site sequences and evolution of spliceosomal introns. Mol Biol Evol 16(11):1528-1534

Lynch M (2007) Genes in pieces. In: The origins of genome architecture. Sunderland, Massachusetts, pp238-270

Niksic M, Romano M, Buratti E, Pagani F, Baralle FE (1999) Functional analysis of *cis*-acting elements regulating the alternative splicing of human CFTR exon 9. Human Mol Genet 8(13):2339-2349 doi: 10.1093/hmg/8.13.2339

Purugganan M, Wessler S (1992) The splicing of transposable elements and its role in intron evolution. Genetica 86:295-303 doi: 10.1007/BF00133728

Qui W, Schisler N, Stoltzfus A (2004) The Evolutionary gain of spliceosomal introns: Sequence and phase preferences. Mol Biol Evol 21(7):1252-1263 doi: 10.1093/molbev/msh120

Robertson HM (1998) Two large families of chemoreceptor genes in nematodes *Caenorhabditis elegans* and *Caenorhabditis briggsae* reveal extensive gene duplication, diversification, movement, and intron loss. Genome Res. 8:449-463 doi: 10.1101/gr.8.5.449

Rogers JH (1989) How were introns inserted into nuclear genes? Trends Genet 5(7):213-216 doi: 10.1016/0168-9525(89)90084-X

Rogozin IB, Lyons-Weiler J, Koonin EV (2000) Intron slidin g in conserved gene families. Genome Analysis 16(10):430-432 doi: 10.1016/S0168-9525(00)02096-5

Rogozin IB, Wolf YI, Sorokin AV, Mirkin BG, Koonin EV (2003) Remarkable interkingdom conservation of intron positions and massive, lineage-specific intron loss and gain in eukaryotic evolution. Current Biol.13:1512-1517 doi: 10.1016/S0960-9822(03)00558-X

Roy SW (2004) The origin of recent introns: transposons? Genome Biol. 2004 5(12)251.1-251.4 doi: 10.1186/gb-2004-5-12-251

Roy SW, Gilbert W (2005) The pattern of intron loss. PNAS 102(3):713-718 doi: 10.1073/pnas.0408274102

Roy SW, Gilbert W (2006) The evolution of spliceosomal introns: patterns, puzzles and progress. Nature Reviews Genetics 7:211-221 doi: 10.1038/nrg1807

Roy SW, Penny D (2006) Smoke without fire: most reported cases of intron gain in nematodes instead reflect intron losses. Mol Biol Evol 23(12)2259-2262 doi: 10.1093/molbev/ms1098

Sakurai A, Fujimori S, Kochiwa H, Kitamura-Abe S, Washio T, Saito R, The RIKEN Genome Exploration Research Group Phase II Team, Carninci P, Hayashizaki Y, Tomita M (2002) On biased distribution of introns in various eukaryotes. Gene 300:89-95 doi: 10.1016/S0378-1119(02)01035-1

Sharp PA (1983) Conversion of RNA to DNA in mammals: Alu-like elements and pseudogenes. Nature 301:471-472 doi: 10.1038/301471a0

Sharp PA (1985) On the origin of RNA splicing and introns. Cell 42:397-400 doi: 10.1016/0092-8674(85)90092-3

Stoltzfus A, Logsdon JM, Palmer JD, Doolittle WF (1997) Intron "sliding" and the diversity of intron positions. PNAS 94:10739-10744 doi: 10.1073/pnas.94.20.10739

Szak ST, Pickeral OK, Makalowski W, Boguski MS, Landsman D, Boeke JD (2002) Molecular archeology of L1 insertions in the human genome. Genome Biol 3(10):research0052-research0052.18 doi: 10.1186/gb-2002-3-10-research0052

Tahira AC, Kubrusly M, Faria MF, Dazzani B, Fonseca RS, Maracaja-Coutinho V, Verjovski-Almeida S, Machado MCC, Reis EM (2011) Long noncoding intronic RNAs are differentially expressed in primary and metastatic pancreatic cancer. Mol Cancer 10:141 doi: 10.1186/1476-4598-10-141

Tarrio R, Ayala FJ, Rodriguez-Trelles F (2008) Alternative splicing: a missing piece in the puzzle of intron gain. PNAS. 105(2):7223-7228 doi: 10.1073/pnas.0802941105

The C. elegans Sequencing Consortium (1998) Genome Sequence of the Nematode C. elegans: A Platform for Investigating Biology. Science 282:2012-2018 doi: 10.1126/science.282.5396.2012

Torriani SF, Stukenbrock EH, Brunner PC, McDonald BA, Croll D (2011) Evidence for extensive recent intron transposition in closely related fungi. Curr Biol 21:2017-2022 doi: 10.1016/j.cub.2011.10.041

Volfovsky N, Haas BJ, Salzberg SL (2003) Computational discovery of internal micro-exons. Genome Res. 13:1216-1221 doi: 10.1101/gr.677503

Zahler AM (2005) Alternative splicing in *C.elgans*. WormBook, ed. http://www.wormbook.org/chapters/www_altsplicing/altsplicing.html Accessed 02 October 2012

APPENDIX

Name of Script	Function
aaPos_Finder.pl	Finds amino acid positions of potential drifted introns
alignAllFamilies.pl	Sends files containing ortholog amino acid sequences to ClustalW2 for alignment
alignOrths.pl	Builds fasta files for each ortholog, sends fasta files to ClustalW2 and then calls cutter.pl to trim alignments
alignOUZinGaps.pl	Uses positional information to insert intron markers after ClustalW2 alignment around gaps
annotateGainLoss.pl	Prints ID information about each intron position for the untrimmed data
blastParse.pl	Reports BLAST hits that are longer than a given cutoff
blastParse2.pl	Parses through M9 blast results and reports back only the first hit
cdsBuilder.pl	Creates a fasta of ortholog CDSs
contigPos_Finder.pl	Finds the contig start position of potential drifted introns
dnaTranslator.pl	Translates a DNA sequence to an AA sequence
driftFinger.pl	Searches for blast hits that contain introns suspected of drift
elegGeneFinder.pl	Compares the CDS length to find longest alternative splicing variant for elegans only
elegOrthologGetter.pl	Finds the orthologs out of the list of longest alt. splicing variant for elegans only
exonBoundryCompiler.pl	Extracts exon-exon junctions for all orthologs
fastCutter.pl	Removes portions of alignments that are ambiguous
fileFiner.pl	Finds files containing intron ID
finisher.pl	Prints GFF line, slightly simplified
gainGeneNamer.pl	Finds the geneIDs that contain gains from alignment files
gainsLosses.pl	Creates an intron position file to identify intron gain/loss from alignments

geneSelector.pl	Searches for geneIDs from a list of other geneIDs
getIntronGainSeq.pl	Uses a list of intron gain IDs to find the intron sequences
inferGainsAndLosses.pl	Uses intron position file to print intron position and gain/loss information
intronAlignmentPos.pl	Rebuilds intron position data from the amino acid alignments after they have gone through fastCutter.pl
intronBlastSorter.pl	Removes hits that overlap original intron position from BLAST M8 output
intronCounter.pl	Counts the number of introns of each gene
intronDashReplacer.pl	Uses modified aligned AA seq and replaces "-" with Xs when the corresponding position is an intron
intronFinder.pl	Finds GTF info of introns for list of geneIDs
intronRelPosFinder.pl	Calculates the intron position relative to the gene
makeFastaByFamily.pl	Makes a fasta file for each ortholog containing the DNA sequence for each taxa
mergeIntPosFiles.pl	Merges all intron position files into one file
orthologGetter.pl	Finds the orthologs out of the list of longest alt. splicing variant
orthologIntronGetter.pl	Creates a fasta of all ortholog intron sequences
ouzReplacer.pl	Puts OUZ into aligned AA sequences
perfectDrift_Finder.pl	Looks for potential drifted introns (using contig start position) among the EST blast 100% matches
postAlignment.pl	Generates gain/loss information and annotates it with intron sequence IDs
preAlignment.pl	Constructs amino acid sequences for each ortholog with inserted intron markers
quickListFinder.pl	Looks for items from one list in another list
reduceGFF.pl	Reduces GFF to intron, exon and cds lines
reduceGFFeleg.pl	Reduces GFF to intron, exon and cds lines for elegans only
reorderTrim.pl	Reorders the sequences in the trimmed alignments to match the order of the untrimmed alignments

secondHit.pl	Parses through blast results and returns only the second best hit along with total number of hits
seqLengthCounter.pl	Calculates the amino acid length of a gene containing gained introns
seqLengthFinder.pl	Makes one file containing the list of ortholog file IDs and the alignment length
seqRandomizer.pl	Scrambles a sequence of characters
species3eliminater.pl	Removes introns or CDSs later found to be missing one of the 4 species in their respective position files.
strandFinder.pl	Adds strand information to intron ID file
trimAll.pl	Sends fasta files to ClustalW2 for alignment and then fastCutter.pl for trimming
trimmedSorter.pl	Makes fasta files of all trimmed sequences for each species.